

Strongly Polynomial 2-Approximations of Discrete Wasserstein Barycenters

Steffen Borgwardt

steffen.borgwardt@ucdenver.edu University of Colorado Denver

Abstract. Wasserstein barycenters correspond to optimal solutions of transportation problems for several marginals. These arise in applications from economics to statistics where an optimal solution over a timeline of data is desired. The theory of barycenters is well-developed when the marginal probability measures are absolutely continuous [1], but computation in this setting is restrictive in all but few special cases.

However, in many applications data is given as a set of probability measures with finite support. First results for this setting revealed several favorable properties of the so-called *discrete barycenters* that arise for discrete data [2]: All barycenters have finite support, and there always is a barycenter for which this support is provably sparse. Further, each barycenter allows for a non-mass splitting optimal transport to each of the discrete marginals. Discrete barycenters can be computed by linear programming, but the sizes of these programs scale exponentially.

In this paper, we discuss approximation algorithms that trade a small error for a significant reduction in computational effort. We begin with a strongly polynomial time 2-approximation of the problem that is based on restricting the possible support of the barycenter to the union of supports of the measures. The resulting measure has sparse support, but its transport to the measures will generally split mass. We then exhibit how to recover the non-mass split property in strongly polynomial time via a local improvement step. This comes at the cost of a larger support.

Finally, we present an iterative scheme that alternates between these two computations. After a finite number of iterations, it terminates with a 2-approximation that retains both favorable properties of an exact barycenter, a sparse support and no mass split at the same time. We conclude with some practical computations. In doing so, we observe an extremely low number of iterations until our toy examples terminate, and a practical approximation error that is vastly lower than 2.

Keywords: discrete barycenter, optimal transport, multiple marginals, 2-approximation, mathematical programming

MSC: 90B80, 90C05, 90C46, 90C90

1 Introduction

Optimal transportation problems for several marginals arise in applications ranging from business and finance over physics to medical imaging and computer vision [3,7,8,10,11,14,15,25,27]. The so-called *Wasserstein barycenters* correspond to optimal solutions to these problems, and have seen much recent attention. See for example [1,2,5,6,9,12,13,18,20,21,23,22,24,28,31,32].

Given probability measures P_1, \dots, P_N on \mathbb{R}^d and a weight vector $\lambda \in \mathbb{R}_{>0}^N$ with $\sum_{i=1}^N \lambda_i = 1$, a (λ -weighted) Wasserstein barycenter is a probability measure \bar{P} on \mathbb{R}^d which satisfies

$$\sum_{i=1}^N \lambda_i W_2(\bar{P}, P_i)^2 = \inf_{P \in \mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^N \lambda_i W_2(P, P_i)^2, \quad (1)$$

where W_2 is the quadratic Wasserstein distance and $\mathcal{P}^2(\mathbb{R}^d)$ is the set of all probability measures on \mathbb{R}^d (with finite second moments). See the monographs [29,30] for a review of the Wasserstein

metric and optimal transportation problems. Further, see [1] for some powerful results, in particular establishing existence, uniqueness and an optimal transport characterization of \bar{P} when P_1, \dots, P_N are continuous and have sufficient regularity.

In many applications, data is given as a set of *discrete probability measures* P_1, \dots, P_N having finite, discrete support in \mathbb{R}^d . A *discrete Wasserstein barycenter* is a probability measure \bar{P} which satisfies (1) for such measures. In [2], some first theoretical results were developed for these discrete barycenters. They mirror those in the continuous case, established in [1], with a few exceptions. First, unlike in the continuous case, there may exist several discrete barycenters for the same set of measures. All of them have discrete support and there always is a discrete barycenter with provably sparse support. Analogously to the continuous case, there always exists a non-mass-splitting optimal transport from a discrete barycenter to each discrete marginal. We would like to note that these results were proven for uniform λ_i in [2], but are readily transferred to the case of a general fixed λ ; see [19].

Discrete barycenters can be computed by linear programming [2,9]. However, these programs scale exponentially in the number of measures N - which gives an incentive to study the possibility of the tradeoff of a small approximation error for a barycenter to obtain a significant reduction in computational effort. In this paper, we discuss such an approach.

First, we present a simple 2-approximation of the problem, based on a restriction of the possible support of the approximate barycenter to the union of supports of the measures, and prove that it runs in strongly polynomial time. The corresponding measure has sparse support but generally does split mass in the optimal transport to each discrete marginal. We then present a local improvement step that recovers the non-mass splitting property and prove that it also runs in strongly polynomial time.

Finally, we use these two efficient algorithms as the building blocks of a finite, iterative algorithm where we alternate between them. The result is a 2-approximation with both sparse support and a non-mass splitting transport at the same time.

In Section 2, we introduce some notation and recall previous related work. In Section 3, we formally explain our main contributions and give a detailed outline of the later sections. In Section 4, we present the necessary proofs. We conclude with a discussion of a practical computation for some toy examples in Section 5.

2 Related Work

We begin by recalling some terminology on discrete barycenters, following [2] and [19]. We are given a set of *discrete probability measures* P_1, \dots, P_N , i.e. the measures all have finite, discrete support $\text{supp}(P_i) \subset \mathbb{R}^d$. We denote the size of the support of P_i as $|P_i| = |\text{supp}(P_i)|$. Further, we are given a weight vector $\lambda \in \mathbb{R}_{>0}^N$ with $\sum_{i=1}^N \lambda_i = 1$. The general definition of a Wasserstein barycenter, as in (1), refers to a probability measure \bar{P} on \mathbb{R}^d which satisfies

$$\sum_{i=1}^N \lambda_i W_2(\bar{P}, P_i)^2 = \inf_{P \in \mathcal{P}^2(\mathbb{R}^d)} \sum_{i=1}^N \lambda_i W_2(P, P_i)^2.$$

For the discrete measures P_1, \dots, P_N , it is not hard to see that all optimizers of (1) must be supported in the finite set $S \subset \mathbb{R}^d$, where

$$S := \left\{ \sum_{i=1}^N \lambda_i x_i : x_i \in \text{supp}(P_i) \right\} \quad (2)$$

is the set of all possible weighted centroids for a combination of support points with one from each measure P_i . In particular, letting $\mathcal{P}_S^2(\mathbb{R}^d) := \{P \in \mathcal{P}^2(\mathbb{R}^d) \mid \text{supp}(P) \subseteq S\}$, the infinite-dimensional problem (1) can be solved by replacing the requirement $P \in \mathcal{P}^2(\mathbb{R}^d)$ with $P \in \mathcal{P}_S^2(\mathbb{R}^d)$ to obtain

$$\phi(\bar{P}) := \inf_{P \in \mathcal{P}_S^2(\mathbb{R}^d)} \sum_{i=1}^N \lambda_i W_2(P, P_i)^2. \quad (3)$$

This yields a finite-dimensional minimization problem, which can be solved by linear programming as follows [2,9,19]:

Let P_1, \dots, P_k be a set of discrete measures and let $\text{supp}(P_i) = \{x_{ik} | k = 1, \dots, |P_i|\}$. Further, let P_0 be another discrete measure and let $\text{supp}(P_0) = \{x_j | j = 1, \dots, |P_0|\}$. Finally, let d_{ik} be the mass of the point x_{ik} in P_i and d_j be the mass of the point x_j in P_0 . Then, we can find the value of $\sum_{i=1}^N \lambda_i W_2(P_0, P_i)^2$, where $\lambda \in \mathbb{R}_{\geq 0}^N$ is a vector of convex weights, by finding the optimal value of the following LP:

$$\begin{aligned}
\min \quad & \sum_{i=1}^N \lambda_i \sum_{j=1}^{|P_0|} \sum_{k=1}^{|P_i|} \|x_j - x_{ik}\|^2 y_{ijk} \\
\sum_{k=1}^{|P_i|} y_{ijk} = & d_j, \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, |P_0|, \\
\sum_{j=1}^{|P_0|} y_{ijk} = & d_{ik}, \quad \forall i = 1, \dots, N, \quad \forall k = 1, \dots, |P_i|, \\
y_{ijk} \geq & 0, \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, |P_0|, \quad \forall k = 1, \dots, |P_i|.
\end{aligned} \tag{4}$$

By introducing variables z_j for the points in a given set $S_0 = \{x_j | j = 1, \dots, |S_0|\}$ to denote the possible mass at $x_j \in S_0$, we obtain an LP that both finds an optimal measure P_0 supported on S_0 , as well as the corresponding transport to get an optimal value for $\sum_{i=1}^N \lambda_i W_2(P_0, P_i)^2$:

$$\begin{aligned}
\min \quad & \sum_{i=1}^N \lambda_i \sum_{j=1}^{|S_0|} \sum_{k=1}^{|P_i|} \|x_j - x_{ik}\|^2 y_{ijk} \\
\sum_{k=1}^{|P_i|} y_{ijk} = & z_j, \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, |S_0|, \\
\sum_{j=1}^{|S_0|} y_{ijk} = & d_{ik}, \quad \forall i = 1, \dots, N, \quad \forall k = 1, \dots, |P_i|, \\
y_{ijk} \geq & 0, \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, |S_0|, \quad \forall k = 1, \dots, |P_i| \\
z_j \geq & 0, \quad \forall j = 1, \dots, |S_0|.
\end{aligned} \tag{5}$$

Note that for P_0 to be a measure, the variables z_j have to satisfy $\sum_{j=1}^{|S_0|} z_j = 1$. But this is a direct consequence of satisfaction of the other constraints. Thus the above program computes a measure represented by

$$(z, y) = ((z_j)_{j=1, \dots, |S_0|}, (y_{ijk})_{i=1, \dots, N, j=1, \dots, |S_0|, k=1, \dots, |P_i|}).$$

By choosing $S_0 = S$, the returned (z, y) , respectively just z , represents a discrete barycenter.

Let us use consider the size of program (5). It consists of $|S_0| + |S_0| \cdot \sum_{i=1}^N |P_i|$ variables and $N \cdot |S_0| + \sum_{i=1}^N |P_i|$ equality constraints. For $S_0 = S$, we get a worst-case bound of $|S_0| = \prod_{i=1}^N |P_i|$. Let now $|P_{\max}| = \max_{i=1, \dots, N} |P_i|$. If all measures have the same number of support points, we get $\sum_{i=1}^N |P_i| = N \cdot |P_{\max}|$ and $\prod_{i=1}^N |P_i| = |P_{\max}|^N$. So we have a linear program of up to $|P_{\max}|^N + |P_{\max}|^N \cdot N \cdot |P_{\max}|$ variables and $N \cdot |P_{\max}|^N + N \cdot |P_{\max}|$ equality constraints.

A more refined analysis reveals that some of the variables and constraints can be redundant. For example, if the different measures overlap in some of their support points, then $|S_0|$ and

consequently the size of the LP become smaller. The example in [2] was computable on a standard laptop, because all measures had the same small support, which had a dramatic effect in reducing $|S_0|$. In general, however, we cannot rule out a scaling of the size of the linear program for $S_0 = S$ that is exponential in N even if $|P_{\max}|$ is fixed, and a polynomial scaling in $|P_{\max}|$ even if N is fixed. This highlights the potential benefit from performing an approximate computation where one reduces the size of $|S_0|$.

The feasible regions of LPs (4) and (5) are bounded, and thus standard arguments of linear programming give us that there is always an optimal vertex. In such a vertex an inclusion-maximal set of variables is set to 0. By a careful analysis of which of the variables z_j, y_{ijk} are equal to 0 in a vertex, it is possible to show a first favorable property: In contrast to the large number of possible support points $|S|$, which can be up to $\prod_{i=1}^N |P_i|$, there always is a barycenter that assigns nonzero weight to less than $\sum_{i=1}^N |P_i|$ of these points. More precisely

Proposition 1. *Let P_1, \dots, P_N be discrete probability measures. Then for any weights $\lambda \in \mathbb{R}_{>0}^n$, there exists a barycenter \bar{P} of these measures such that*

$$|\bar{P}| \leq \sum_{i=1}^N |P_i| - N + 1. \quad (6)$$

In fact, a proof of this claim (see Theorem 2 in [2], Theorem 19 in [19]) is only based on being at the vertex of the underlying polyhedron. It does not require this vertex to be a barycenter (but note there always is a vertex that corresponds to a barycenter). Further, the argument also works if a support set $S_0 \neq S$ is used. The LP (5) optimizes the objective function (1) over the set $\mathcal{P}_{S_0}^2(\mathbb{R}^d)$ of all probability measures P with support in S_0 , but in general there may not be a barycenter supported in $S_0 \neq S$. We call the optimal measure in $\mathcal{P}_{S_0}^2(\mathbb{R}^d)$ an S_0 -barycenter, an *approximation of the barycenter in S_0* , or when the context is clear simply an *approximate barycenter*. For these different support sets, we have the following direct extension of Theorem 1.

Corollary 1. *Let P_1, \dots, P_N be discrete probability measures in \mathbb{R}^d , let $S_0 = \{x_j : j = 1, \dots, |S_0|\} \subset \mathbb{R}^d$, and let $\mathcal{P}_{S_0}^2(\mathbb{R}^d)$ be the set of all probability measures P with support in S_0 . Then for any weights $\lambda \in \mathbb{R}_{>0}^n$, there exists an approximation \bar{P}_0 of the barycenter in S_0 such that*

$$|\bar{P}_0| \leq \sum_{i=1}^N |P_i| - N + 1. \quad (7)$$

Second, for a fixed barycenter \bar{P} , it is possible to show the existence of a *non-mass splitting* transport from each of the support points of the barycenter to the measures: This means that for all $x_j \in \text{supp}(\bar{P})$ with mass d_j and for each $i \in \{1, \dots, N\}$, there is exactly one k with $y_{ijk} = d_j$ for the corresponding variables in LP (4), while for all $k' \neq k$ we have $y_{ijk'} = 0$. So each support point of a barycenter only transports mass to exactly one support point in each measure.

To see this, note that the barycenter only consists of support points which are the centroids (weighted according to λ) of a set of support points in the measures, one for each measure. (See Theorem 1 in [2], Theorem 18 in [19]). We informally say that a support point does not split mass or that a barycenter is non-mass splitting if the property holds for all support points. With this wording, we can formally state the above as

Proposition 2. *Let P_1, \dots, P_N be discrete probability measures, and let \bar{P} be a barycenter for these measures. Then \bar{P} is non-mass splitting.*

3 Main Results

In this paper, we study an approximation algorithm for the barycenter problem where we reduce the size of S_0 . This is motivated by the unfavorable scaling of LP (5) with respect to $|S_0|$; see the discussion in Section 2. We here outline our main results, Section 4 provides the necessary proofs and some examples.

3.1 A strongly polynomial 2-approximation

Recall that the set of possible support points of a barycenter is

$$S := \left\{ \sum_{i=1}^N \lambda_i x_i : x_i \in \text{supp}(P_i) \right\}, \quad (8)$$

which may consist of up to $\prod_{i=1}^N |P_i|$ points. This is a much larger number than the size of the union of supports of the measures

$$S_{\text{org}} := \bigcup_{i=1}^N \text{supp}(P_i), \quad (9)$$

which satisfies $|S_{\text{org}}| \leq \sum_{i=1}^N |P_i|$ with equality if and only if the supports are disjoint. Note that the maximal size of S_{org} only barely exceeds the bound in Proposition 6. This could lead to an intuition that an approximation of the barycenter via a measure with support in S_{org} could be quite bad. But just the opposite is true:

Our first result is that by setting S_0 to S_{org} in program (5), i.e. by performing the search for an optimizer of the barycenter objective function over the set $\mathcal{P}_{S_{\text{org}}}^2(\mathbb{R}^d)$, or shorter $\mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$, of all probability measures P_0 with $\text{supp}(P_0) \subset S_0 = S_{\text{org}}$ gives a 2-approximation for the original problem. This bound is tight.

Theorem 1. *Let \bar{P} be a barycenter and let \bar{P}_{org} be a solver for*

$$\phi(\bar{P}_{\text{org}}) := \sum_{i=1}^N \lambda_i W_2(\bar{P}_{\text{org}}, P_i) = \inf_{P_0 \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^d)} \sum_{i=1}^N \lambda_i W_2(P_0, P_i)^2. \quad (10)$$

Then

$$\phi(\bar{P}_{\text{org}}) \leq 2 \cdot \phi(\bar{P})$$

and this bound can become tight, i.e. there is a set of measure P_1, \dots, P_N and a set of weights $\lambda_1, \dots, \lambda_N$ for which $\phi(\bar{P}_{\text{org}}) = 2 \cdot \phi(\bar{P})$.

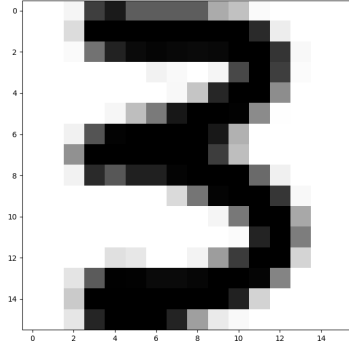
We formally denote the choice of S_{org} in program (5), as performed for Theorem 1, as Algorithm 1. Let us highlight the difference between the support for an exact barycenter and for this approximation using Figure 1:

The first two rows of the figure show four handwritten digits scanned into a 16×16 grid. (See [16] for some information on this data set.) These are the measures P_1, \dots, P_4 . The varying shades of grey indicate different masses at the support points of the grid – the darker, the larger the mass. The masses for each measure add up to 1. The bottom row depicts an exact barycenter and a 2-approximation in the original 16×16 grid (for all $\lambda_i = \frac{1}{4}$). Note that the support grid for the exact barycenter is 64×64 .

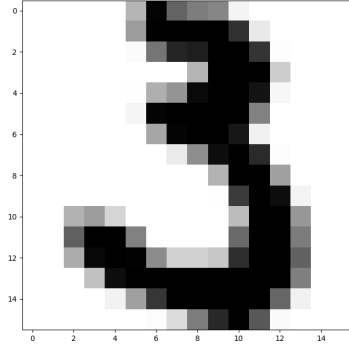
An approximate barycenter as computed in Algorithm 1 satisfies the sparsity condition stated in Corollary 1, so in particular there always is a sparse 2-approximate barycenter supported in S_{org} . However, it is not hard to give an example with support points that split mass, in contrast to Proposition 2 (and we do so in Section 4.1). We close our discussion of the algorithm by identifying the favorable running time of the algorithm.

Theorem 2. *For all rational input, a 2-approximate barycenter can be computed in strongly polynomial time.*

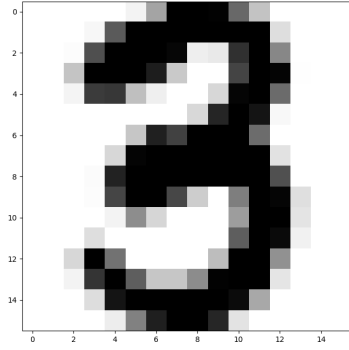
A proof is based on exhibiting the existence of a linear program for this computation, where the numbers of the linear program can be computed in strongly polynomial time and are of strongly polynomial size. Strongly polynomial solvability of this program then follows from the constraint matrix only having entries in $\{-1, 0, 1\}$ – the form of numbers in the objective function and right-hand sides does not matter [26].



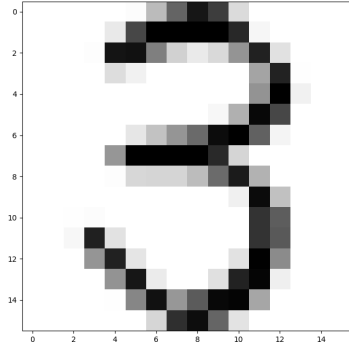
Measure P_1



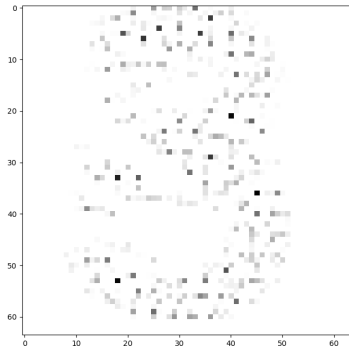
Measure P_2



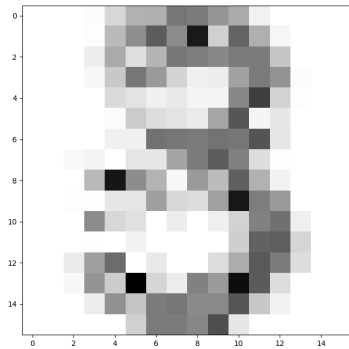
Measure P_3



Measure P_4



Barycenter \bar{P}



Approximate Barycenter \bar{P}_{org}

Fig. 1: Four measures P_1, \dots, P_4 supported on a 16×16 grid in the first two rows. The bottom row shows a barycenter \bar{P} and an approximate barycenter \bar{P}_{org} . Note that \bar{P} is supported in a 64×64 grid, while the support of \bar{P}_{org} lies in the original 16×16 grid.

Algorithm 1 A 2-approximate barycenter in the original support

Input

- Probability measures $P_1, \dots, P_N \subset \mathbb{R}^d$
- $\lambda_1, \dots, \lambda_N > 0$ with $\sum_{i=1}^N \lambda_i = 1$

Algorithm

Compute an approximate barycenter \bar{P}_{org} in S_{org} by solving

$$\begin{aligned} \min \quad & \phi(\bar{P}_{\text{org}}) := \sum_{i=1}^N \lambda_i \sum_{j=1}^{|S_{\text{org}}|} \sum_{k=1}^{|P_i|} \|x_j - x_{ik}\|^2 y_{ijk} \\ & \sum_{k=1}^{|P_i|} y_{ijk} = z_j, \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, |S_{\text{org}}|, \\ & \sum_{j=1}^{|S_{\text{org}}|} y_{ijk} = d_{ik}, \quad \forall i = 1, \dots, N, \quad \forall k = 1, \dots, |P_i|, \\ & y_{ijk} \geq 0, \quad \forall i = 1, \dots, N, \quad \forall j = 1, \dots, |S_{\text{org}}|, \quad \forall k = 1, \dots, |P_i| \\ & z_j \geq 0, \quad \forall j = 1, \dots, |S_{\text{org}}| \end{aligned}$$

and return (z, y) to represent \bar{P}_{org} .

3.2 Recovery of non-Mass Split

We then present an algorithm that begins with a 2-approximation computed by Algorithm 1 and recovers the non-mass split property, while keeping or improving the approximation error. The algorithm greedily breaks up each support point (that still splits mass) of the approximate barycenter into several non-mass splitting support points. In the end, all of these support points are combined to a new approximate barycenter. Algorithm 2 sums up the approach in pseudocode.

In Step 1, the approximate barycenter \bar{P}_{org} is broken up into disjoint parts – each part corresponds to a support point $s_l = x_{t_l}$ in the approximate barycenter. By construction, each P_i^l consists of those support points in P_i to which s_l transports mass. The mass of a support point in P_i^l equals the mass it receives as transport from s_l . In the end of the step, we index the support points in P_i^l and their masses, so we do not have to refer to z or y in the subsequent steps.

Steps 2 and 3 both are based on the construction of so-called *lexicographically maximal* vectors. We call a vector $a = (a_1, \dots, a_n)$ lexicographically larger than a vector $b = (b_1, \dots, b_n)$ if there is an index $j \leq n$ such that $\sum_{i=1}^j a_i > \sum_{i=1}^j b_i$, while $\sum_{i=1}^l a_i \geq \sum_{i=1}^l b_i$ for all $l < j$. For example, the vector $a = (2, 2, 0, 1)$ is lexicographically larger than $b = (2, 1, 5, 10)$. Lexicographic maximality with respect to a set S then states that there is no lexicographically larger vector in S . Note that the term gives rise to a total ordering.

Step 2 transforms (d_1, \dots, d_r) to be lexicographically maximal for all approximate barycenters supported in $\text{supp}(\bar{P}_{\text{org}}) = \{s_1, \dots, s_r\}$. As \bar{P}_{org} is optimal output from Algorithm 1, it is in particular optimal over $\text{supp}(\bar{P}_{\text{org}})$. Informally, among all approximate barycenters over $\text{supp}(\bar{P}_{\text{org}})$, as much mass as possible is moved to support points with lowest indices.

The two loops for l and j establish an order for checking whether mass can be moved from s_l to s_j . The indices q_i selected in a) identify support points in the P_i^l that lie the furthest in direction of $s_j - s_l$. Their weighted centroid c is a maximizer of $\|c - s_l\|^2 - \|c - s_j\|^2$. This difference is bounded above by 0 because of optimality of \bar{P}_{org} . However, if $\|c - s_l\|^2 = \|c - s_j\|^2$, which is checked in b), then mass can be shifted from s_l to s_j to make (d_1, \dots, d_r) lexicographically larger, while keeping optimality over $\text{supp}(\bar{P}_{\text{org}})$. The remainder of b) is a technical description of this shift of mass.

In Step 3, we then perform a greedy routine to spread out the mass of each s_l to a set of support points that do not split mass anymore. We do so by picking a set of lexicographically maximal

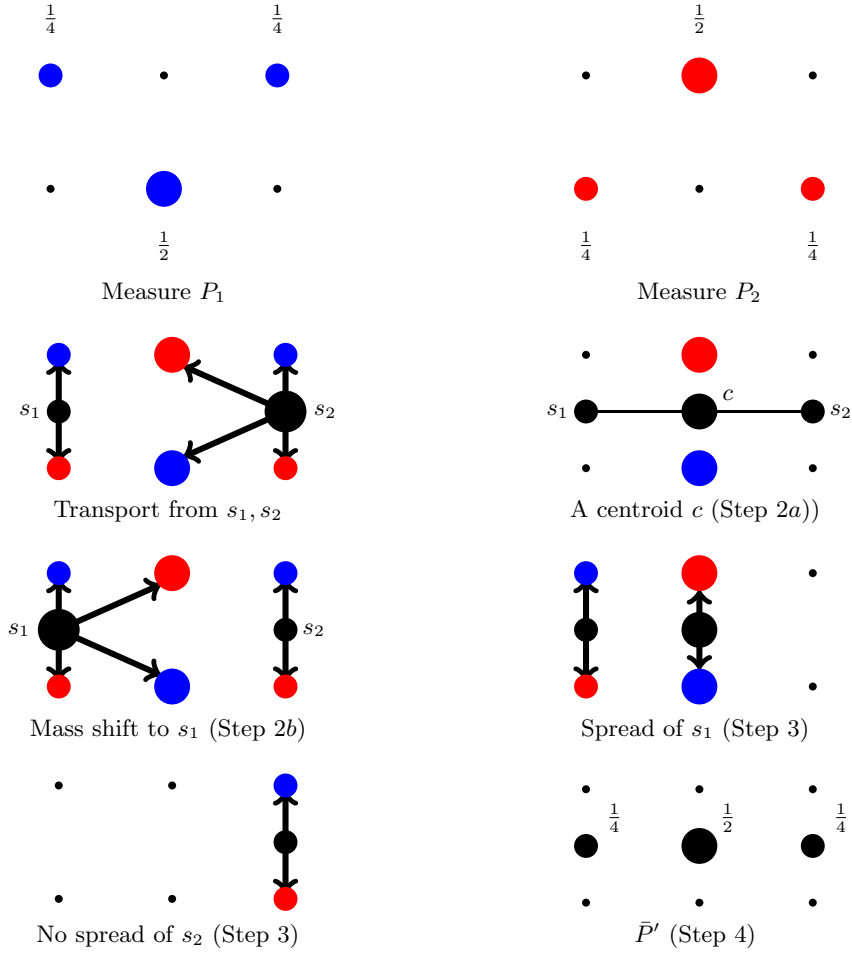


Fig. 2: Two measures P_1, P_2 in the top row and a run of Steps 2 – 4 of Algorithm 2 for given support points s_1, s_2 of mass $d_1 = \frac{1}{4}, d_2 = \frac{3}{4}$.

Algorithm 2 Recovery of non-mass split

Input

- Probability measures $P_1, \dots, P_N \subset \mathbb{R}^d$
- 2-approximate barycenter \bar{P}_{org} represented by (z, y) computed by Algorithm 1
- $\lambda_1, \dots, \lambda_N > 0$ with $\sum_{i=1}^N \lambda_i = 1$

Algorithm
1. (Break up \bar{P}_{org} into parts for each support point)

Let $\text{supp}(\bar{P}_{\text{org}}) = \{s_1, \dots, s_r\} = \{x_{t_1}, \dots, x_{t_r}\}$ with corresponding masses $d_1 = z_{t_1}, \dots, d_r = z_{t_r} > 0$.
 For each $l \leq r$ and $i \leq N$, construct P_i^l (a set of support points with masses) by the rule:

if $y_{it_l k} > 0$ to $x_{ik} \in \text{supp}(P_i)$, then add x_{ik} to $\text{supp}(P_i^l)$ with mass $y_{it_l k}$

Now index $P_i^l = \{x_{i1}^l, \dots, x_{i|P_i^l|}^l\}$ with corresponding masses $d_{i1}^l, \dots, d_{i|P_i^l|}^l$ for all $l \leq r$ and $i \leq N$.

2. (Make (d_1, \dots, d_r) lexicographically maximal)

For $l = r$ descending to $l = 1$

For $j = 1$ ascending to $j = l - 1$

a) For each $i \leq N$, identify the index $q_i = \arg \max_{q \leq |P_i^l|} (s_j - s_l)^T x_{iq}^l$. Then compute the

weighted centroid $c = \sum_{i=1}^N \lambda_i x_{iq_i}^l$ from the corresponding support points.

b) If $\|c - s_j\|^2 = \|c - s_l\|^2$ then

Identify the minimal mass $d_{\min} = \min_{i \leq N} d_{iq_i}^l$ among the $x_{iq_i}^l$.

Set $d_l = d_l - d_{\min}$ and $d_{iq_i}^l = d_{iq_i}^l - d_{\min}$ for all $i \leq N$.

For all $i \leq N$, if $d_{iq_i}^l = 0$, remove $x_{iq_i}^l$ from $\text{supp}(P_i^l)$ and reindex P_i^l and $d_{i1}^l, \dots, d_{i|P_i^l|}^l$.

For all $i \leq N$, add $x_{iq_i}^l$ to $\text{supp}(P_i^j)$ if it is not in it yet. In this case, $|P_i^j|$ increases by one and we index the support point as $x_{i|P_i^j|}^j$ (with $d_{i|P_i^j|}^j = 0$).

Let now p_i be such that $x_{ip_i}^j = x_{iq_i}^l$ for all $i \leq N$.

Set $d_j = d_j + d_{\min}$ and $d_{ip_i}^j = d_{ip_i}^j + d_{\min}$ for all $i \leq N$.

If $d_l > 0$, go back to a).

3. (Spread out each support point to a set of weighted centroids)

For $l = 1$ ascending to $l = r$

Create an empty \bar{P}^l (a set of support points with masses).

a) For each $i \leq N$, pick a lexicographically maximal support point $x_{iq_i}^l$ from P_i^l . Then compute

the weighted centroid $c = \sum_{i=1}^N \lambda_i x_{iq_i}^l$.

b) Identify the minimal mass $d_{\min} = \min_{i \leq N} d_{iq_i}^l$ among the $x_{iq_i}^l$.

Set $d_l = d_l - d_{\min}$ and $d_{iq_i}^l = d_{iq_i}^l - d_{\min}$ for all $i \leq N$.

For all $i \leq N$, if $d_{iq_i}^l = 0$, remove $x_{iq_i}^l$ from $\text{supp}(P_i^l)$ and reindex P_i^l and $d_{i1}^l, \dots, d_{i|P_i^l|}^l$.

Add c to $\text{supp}(\bar{P}^l)$ with mass d_{\min} .

If $d_l > 0$, go back to a).

4. (Combine a new approximate barycenter)

Combine the \bar{P}^l to an approximate barycenter $\bar{P}' = \sum_{l=1}^r \bar{P}^l$. Return \bar{P}' .

support points $x_{iq_i}^l$ in each P_i^l (i.e., we pick an $x_{iq_i}^l$ with a largest first coordinate, and among those one with a largest second coordinate, and so on). Then we move mass d_{\min} to the weighted centroid $c = \sum_{i=1}^N \lambda_i x_{iq_i}^l$, where d_{\min} is the minimal mass among the $d_{q_i}^l$. We repeat this scheme until all the mass of a support point has been spread out, then continue with the next support point.

Finally, in Step 4 we combine the results of Step 3 to a new approximate barycenter. It is at least as good an approximation of a true barycenter as \bar{P}_{org} . This is because in Step 3, for any chosen set of support points $x_{iq_i}^l$ we put the corresponding mass on their weighted centroid, which is a best-possible choice (and certainly at least as good as transport from s_l).

The weighted centroids created for a single s_l in Step 3 will all differ from each other (due to the choice of lexicographically maximal $x_{iq_i}^l$) and only transport to a single support point in $\text{supp}(P_i^l) \subset \text{supp}(P_i)$ by construction. By the shift of mass in Step 2, the weighted centroids created for different s_j, s_l are always different from each other. Together, these are the reasons why the output of Algorithm 2 is non-mass splitting.

Let us discuss a small example for Steps 2 – 4 of the algorithm.

Example 1. Consider the two measures P_1, P_2 depicted at the top of Figure 2 and let $\lambda_1 = \lambda_2 = \frac{1}{2}$. The volume of the filled circles represents the corresponding mass. They receive their mass transported from two fixed support points s_1, s_2 of mass $d_1 = \frac{1}{4}, d_2 = \frac{3}{4}$ (second row, left). (Note that $s_1, s_2 \notin S_{\text{org}}$, but this does not matter for this example.)

The two center points, which receive their mass from s_2 , have a centroid c that is equally far from s_1 and s_2 (second row, right). These two points would be selected in Step 2a) of Algorithm 2 and their mass shifted to s_1 from s_2 in Step 2b): Now $d_1 = \frac{3}{4}, d_2 = \frac{1}{4}$ (third row, left).

In Step 3, the mass of s_1 and s_2 is spread out to a set of centroids that transport to just a single support point in each measure. The result for s_1 is depicted in the third row (left). By lexicographically maximal choice of the support points, the central point of mass $\frac{1}{2}$ is constructed first, followed by the left one of mass $\frac{1}{4}$. s_2 is not changed, because it already is the centroid of a set of single support points in each measure (fourth row, left).

These measures are combined to form \bar{P}' in Step 4 (fourth row, right) and the algorithm stops. In this example, we actually found an exact barycenter. \square

We sum up the favorable properties of the algorithm in Theorem 3. A detailed proof is given in Section 4.2.

Theorem 3. *Algorithm 2 returns a measure \bar{P}' supported on a subset of S for which $\phi(\bar{P}') \leq 2 \cdot \phi(\bar{P})$, where \bar{P} is a barycenter. Further $|\bar{P}'| \leq (\sum_{i=1}^N |P_i| - N + 1)^2$ and \bar{P}' is non-mass splitting.*

Again, we close our discussion of the algorithm by identifying its strongly polynomial running time.

Theorem 4. *For all rational input, a 2-approximate non-mass splitting barycenter can be computed in strongly polynomial time.*

3.3 An Iterative Local Improvement

Finally, we combine Algorithms 1 and 2 to an iterative scheme, which is denoted as Algorithm 3.

The algorithm begins by computing an approximate barycenter in S_{org} , as in Algorithm 1. Then Algorithm 2 is used to spread out its support points to non-mass split, which also improves the approximation error. The result is a new measure with support S_0 . We set $S_{\text{org}} = S_0$ and repeat Algorithm 1 to find an optimal approximate barycenter over this support. Then its support points are spread out again. This scheme is repeated until there is no improvement anymore.

After a finite number of iterations, the algorithm terminates with a sparse approximate barycenter supported in S that satisfies the non-mass split condition, i.e. an approximation that mirrors both favorable properties of a true barycenter.

Theorem 5. *Algorithm 3 returns a measure \bar{P}' supported on a subset of S for which $\phi(\bar{P}') \leq 2 \cdot \phi(\bar{P})$, where \bar{P} is a barycenter. Further $|\bar{P}'| \leq \sum_{i=1}^N |P_i| - N + 1$ and \bar{P}' is non-mass splitting.*

Algorithm 3 Iterative local improvement

Input

- Probability measures $P_1, \dots, P_N \subset \mathbb{R}^d$
- $\lambda_1, \dots, \lambda_N > 0$ with $\sum_{i=1}^N \lambda_i = 1$

Algorithm

1. Compute an approximate barycenter \bar{P}_{org} in S_{org} using Algorithm 1.
 2. Use \bar{P}_{org} as input for Algorithm 2 to find an improved approximate barycenter \bar{P}' supported in S' .
If $\phi(\bar{P}') < \phi(\bar{P}_{\text{org}})$, set $S_{\text{org}} = S'$ and go back to 1. Else return \bar{P}' .
-

We prove Theorem 5 in Section 4.3 and discuss an example that highlights a (peculiar) scenario, where the algorithm does not improve on the initial solution, but terminates straight away. In Section 5 we conclude the paper by highlighting some observations about practical computations using our algorithms.

4 Proofs

4.1 Proofs for 3.1

We begin by proving Theorem 1, and in doing so proving the correctness of Algorithm 1.

Theorem 1 *Let \bar{P} be a barycenter and let \bar{P}_{org} be a solver for*

$$\phi(\bar{P}_{\text{org}}) := \sum_{i=1}^N \lambda_i W_2(\bar{P}_{\text{org}}, P_i) = \inf_{P_0 \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^d)} \sum_{i=1}^N \lambda_i W_2(P_0, P_i)^2. \quad (11)$$

Then

$$\phi(\bar{P}_{\text{org}}) \leq 2 \cdot \phi(\bar{P})$$

and this bound can become tight, i.e. there is a set of measure P_1, \dots, P_N and a set of weights $\lambda_1, \dots, \lambda_N$ for which $\phi(\bar{P}_{\text{org}}) = 2 \cdot \phi(\bar{P})$.

Proof. Recall that $\text{supp}(\bar{P}) \subset S$. A barycenter \bar{P} consists of a set of support points c in S . We denote the mass of a support point c by d_c . By Proposition 1, each support point c transports its mass to exactly one support point x_i in each P_i for all $i \leq N$. Due to optimality of \bar{P} , c is the weighted centroid $c = \sum_{i=1}^N \lambda_i x_i$ of these points. This can be seen by

$$\begin{aligned} \sum_{i=1}^N \lambda_i \|(s+c) - x_i\|^2 &= (s+c)^T(s+c) - 2(s+c)^T c + \sum_{i=1}^N \lambda_i x_i^T x_i = \\ &= (s^T s + 2s^T c + c^T c) - 2s^T c - 2c^T c + \sum_{i=1}^N \lambda_i x_i^T x_i = s^T s - c^T c + \sum_{i=1}^N \lambda_i x_i^T x_i, \end{aligned}$$

which is minimal for $s^T s = 0$, i.e. $s = 0$.

Each support point c contributes $d_c \cdot \sum_{i=1}^N \lambda_i \|c - x_i\|^2$ to the corresponding value $\phi(\bar{P})$. Let now

$s \in S_{\text{org}} = \bigcup_{i=1}^N \text{supp}(P_i)$ be such that $\|s - c\|^2$ is minimal and note that

$$\begin{aligned} \sum_{i=1}^N \lambda_i \|s - x_i\|^2 &= s^T s - 2c^T s + \sum_{i=1}^N \lambda_i x_i^T x_i = (s^T s - 2c^T s + c^T c) + \\ &+ (c^T c - 2c^T c + \sum_{i=1}^N \lambda_i x_i^T x_i) = \sum_{i=1}^N \lambda_i (\|s - c\|^2 + \|c - x_i\|^2) \end{aligned}$$

for any s . By choice of s and the fact that $x_i \in \text{supp}(P_i)$, we know $\|s - c\|^2 \leq \|c - x_i\|^2$ for all $i \leq N$, so we get

$$\sum_{i=1}^N \lambda_i \|s - x_i\|^2 = \sum_{i=1}^N \lambda_i (\|s - c\|^2 + \|c - x_i\|^2) \leq 2 \cdot \sum_{i=1}^N \lambda_i \|c - x_i\|^2.$$

Thus the transport from s , instead of from c itself, introduces an approximation error of 2, i.e. each such s contributes at most $2 \cdot d_c \sum_{i=1}^N \lambda_i \|c - x_i\|^2$ to the corresponding value $\phi(\bar{P}_{\text{org}})$.

As this argument holds for all of the weighted centroids $c \in \text{supp}(\bar{P})$ and corresponding closest $s \in S_{\text{org}}$, this shows the existence of a probability measure $\bar{P}_{\text{org}} \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$ with approximation error 2 with respect to ϕ .

It remains to prove that the bound can be tight. We do so by exhibiting a small example. Let P_1, P_2 be two measures with just a single support point $x_1 \in \text{supp}(P_1)$, $x_2 \in \text{supp}(P_2)$, each of mass 1. Then \bar{P} consists of the single support point $c = \lambda_1 x_1 + \lambda_2 x_2$ of weight 1 and thus

$$\begin{aligned} \phi(\bar{P}) &= \lambda_1 \cdot \|c - x_1\|^2 + \lambda_2 \cdot \|c - x_2\|^2 = \lambda_1 \cdot \|(\lambda_1 - 1)x_1 + \lambda_2 x_2\|^2 + \lambda_2 \cdot \|\lambda_1 x_1 + (\lambda_2 - 1)x_2\|^2 = \\ &= \lambda_1 \cdot \|\lambda_2(x_2 - x_1)\|^2 + \lambda_2 \cdot \|\lambda_1(x_1 - x_2)\|^2 = \lambda_1 \lambda_2 (\lambda_2 + \lambda_1) \|x_2 - x_1\|^2 = \lambda_1 \lambda_2 \|x_2 - x_1\|^2. \end{aligned}$$

In contrast, the restriction of an approximate barycenter \bar{P}_{org} to possible support $S_{\text{org}} = \{x_1, x_2\}$ would give $\phi(\bar{P}_0) = \min\{\lambda_1, \lambda_2\} \cdot \|x_2 - x_1\|^2$. Note $\lambda_1 \cdot \lambda_2 \geq \frac{1}{2} \min\{\lambda_1, \lambda_2\}$, with equality if and only if $\lambda_1 = \lambda_2 = \frac{1}{2}$. In this case, $\phi(\bar{P}_{\text{org}}) = 2 \cdot \phi(\bar{P})$. \square

In general $\lambda_i = \frac{1}{N}$ does not imply that the approximation error is tight. For example, let P_1, P_2, P_3 be three measures with single support points x_1, x_2 , and $x_3 = \frac{1}{2}(x_1 + x_2)$. Then the single weighted centroid c satisfies $c = x_3 \in \text{supp}(P_3) \subset S_{\text{org}}$. This implies $\phi(\bar{P}_{\text{org}}) = \phi(\bar{P})$.

But even for $\text{supp}(\bar{P}) \not\subset S_{\text{org}}$, $\lambda_i = \frac{1}{N}$ does not imply that the approximation error is tight. Let $s \in S_{\text{org}}$ be such that $\|s - c\|^2$ is minimal for a given weighted centroid $c \notin S_{\text{org}}$ transporting to x_1, \dots, x_N with $x_i \in P_i$. Then the approximation error 2 is not tight if $\|c - x_i\|^2 \neq \|c - x_j\|^2$ for any $i \neq j$. This is because then

$$\sum_{i=1}^N \lambda_i \|s - x_i\|^2 = \sum_{i=1}^N \lambda_i (\|c - x_i\|^2 + \|s - c\|^2) < 2 \cdot \sum_{i=1}^N \lambda_i \|c - x_i\|^2,$$

as there must be a $j \leq N$ with $\|c - x_j\|^2 > \|s - c\|^2$.

While \bar{P}_{org} is guaranteed to have sparse support by Corollary 1, here is an example for a split of mass in the transport.

Example 2. We revisit the measures used for Example 1. Measure $P_{\text{org}} \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$ is one of several optimal barycenter approximations in the original support set S_{org} . It only consists of two support points, while both measures have three support points. Thus, there must be a support points that transports mass to more than one support point in the same P_i . It is depicted next to the measure: The top support point transports to all the support points in the upper half of the layout (including mass $\frac{1}{2}$ to the red support point to which it is identical), the bottom support point transports to all the support points in the lower half. This is a split of mass transport that does not happen for a true barycenter, which is depicted in the bottom of the figure. \square

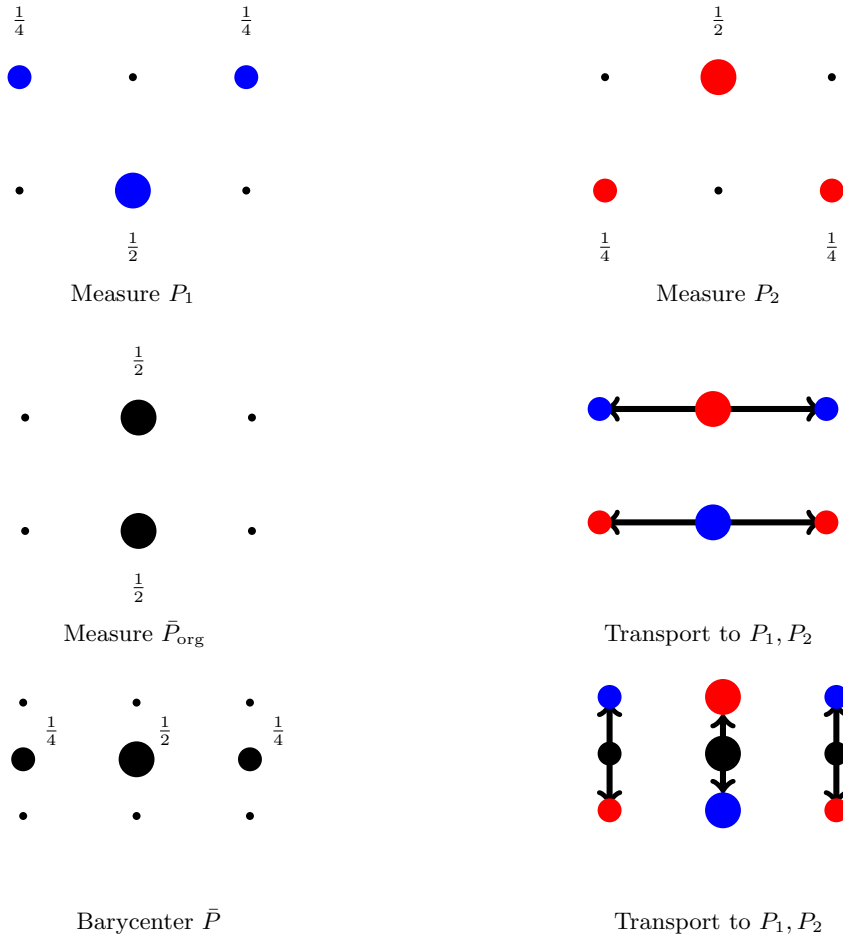


Fig. 3: Two measures P_1, P_2 in the top row. An optimal approximate barycenter $\bar{P}_{\text{org}} \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^2)$ and the corresponding mass splitting transport in the second row. The true barycenter and a corresponding non-mass splitting transport in the third row.

Next, we take a closer look on the linear program that is solved to find \bar{P}_{org} as in Theorem 1. We do so to prove that it can be solved in strongly polynomial time.

Theorem 2 *For all rational input, a 2-approximate barycenter can be computed in strongly polynomial time.*

Proof. Recall that linear programs are generally weakly polynomial-time solvable, i.e. the number of arithmetic operations necessary to solve them depends polynomially on the number of variables and constraints and polynomially on the absolute of numbers in the input. However, it suffices to restrict this dependency to only the absolute of numbers in the constraint matrix – the numbers in the objective function or the right-hand side do not matter [26].

Note that the constraint matrix of program (5) for S_{org} only consists of entries in $\{-1, 0, 1\}$. For the claim of strongly polynomial solvability, this means we only have to prove that the number of variables and constraints of the program is strongly polynomial in the size of the input, and that the numbers that appear in the objective function and right hand side can be computed from the original input in strongly polynomial time.

So let \mathcal{I} be an instance of the problem and $|\mathcal{I}|$ the number of bits to represent the input. First, note that any representation of the input \mathcal{I} has to satisfy $|\mathcal{I}| \geq \sum_{i=1}^N |P_i| \geq N$. As $|S_0| = |S_{\text{org}}| \leq \sum_{i=1}^N |P_i| \leq |\mathcal{I}|$, we see that program (5) has a strongly polynomial number of constraints and variables.

The actual numbers that appear in the program are of types λ_i , d_{ik} , or $\|x_j - x_{ik}\|^2$. The λ_i and d_{ik} appear directly in the input, and so do x_j and x_{ik} . As we use rational input, $\|x_j - x_{ik}\|^2 = (x_j - x_{ik})^T (x_j - x_{ik})$ is a rational number derived by the sum over products of pairs of coefficients in x_j and x_{ik} . This implies that $\|x_j - x_{ik}\|^2$ can be computed in strongly polynomial time (polynomial in $\log x_j + \log x_{ik}$), as well as represented with a number of bits that is strongly polynomial in the number of bits of the original representation of x_j, x_{ik} . \square

4.2 Proofs for 3.2

We begin by proving Theorem 3.

Theorem 3 *Algorithm 2 returns a measure \bar{P}' supported on a subset of S for which $\phi(\bar{P}') \leq 2 \cdot \phi(\bar{P})$, where \bar{P} is a barycenter. Further $|\bar{P}'| \leq (\sum_{i=1}^N |P_i| - N + 1)^2$ and \bar{P}' is non-mass splitting.*

Proof. First, note that the P_i^l constructed in Step 1 satisfy $\text{supp}(P_i^l) \subset \text{supp}(P_i)$. Thus $\text{supp}(\bar{P}^l) \subset S$, and consequently $\text{supp}(\bar{P}') \subset S$. Further, $\bar{P}' = \sum_{l=1}^r \bar{P}^l$ is a measure. This is because $\sum_{l=1}^r d_l = \sum_{l=1}^r z_{t_l} = 1$, Step 2 only changes d to be lexicographically maximal without changing the total sum, and the total mass in \bar{P}^l equals d_l by construction. So \bar{P}' is a measure supported in S .

Second, we prove correctness of Step 2. We show that a lexicographically maximal (d_1, \dots, d_r) among all approximate barycenters in $\text{supp}(P_{\text{org}})$ is created. Further, we show that the objective function value $\phi(\bar{P}_{\text{org}})$ does not change during the shift of mass. For a simple wording, let \bar{P}_{lex} be the measure corresponding to (d_1, \dots, d_r) after Step 2. So we will prove $\phi(\bar{P}_{\text{org}}) = \phi(\bar{P}_{\text{lex}})$.

Let $x_{iq_i}^l \in P_i^l$ for $i \leq N$ and $c = \sum_{i=1}^N \lambda_i x_{iq_i}^l$, as in Step 2a). Then $\|c - s_l\| \leq \|c - s_j\|$ for all $j \neq l$.

To see this, recall

$$\sum_{i=1}^N \lambda_i \|s - x_{iq_i}^l\|^2 = \sum_{i=1}^N \lambda_i (\|s - c\|^2 + \|c - x_{iq_i}^l\|^2),$$

as demonstrated in the proof of Theorem 1. If $\|c - s_l\| > \|c - s_j\|$, \bar{P}_{org} would not have been optimal.

By picking $q_i = \arg \max_{q \leq |P_i^l|} (s_j - s_l)^T x_{iq}^l$ in Step 2a), we pick the $x_{iq_i}^l$ such that for their weighted centroid $c = \sum_{i=1}^N \lambda_i x_{iq_i}^l$ the difference $\|c - s_l\|^2 - \|c - s_j\|^2 \leq 0$ is maximized. Only if $\|c - s_l\|^2 = \|c - s_j\|^2$, mass is shifted from s_l to s_j . But this implies that the approximation error does not change, because then

$$\sum_{i=1}^N \lambda_i \|s_j - x_{iq_i}^l\|^2 = \sum_{i=1}^N \lambda_i (\|s_j - c\|^2 + \|c - x_{iq_i}^l\|^2) = \sum_{i=1}^N \lambda_i (\|s_l - c\|^2 + \|c - x_{iq_i}^l\|^2) = \sum_{i=1}^N \lambda_i \|s_l - x_{iq_i}^l\|^2.$$

So we obtain $\phi(\bar{P}_{\text{org}}) = \phi(\bar{P}_{\text{ex}})$.

By definition of the running indices l and j , mass can only be moved from support points of larger index l to support points of smaller index i . For each pair of l and j , we repeat this shift of mass until there is no weighted centroid with $\|c - s_l\| = \|c - s_j\|$ anymore. Due to decreasing l in the outer loop and increasing j in the inner loop, (d_1, \dots, d_r) is transformed to be lexicographically maximal for all approximate barycenters with support contained in $\text{supp}(\bar{P}_{\text{org}})$.

Next, we prove correctness of Steps 3 and 4. We show that the \bar{P}^l are non-mass splitting and that $\phi(\bar{P}_{\text{ex}}) \geq \phi(\bar{P}')$. Further, we show that $|\bar{P}'| \leq (\sum_{i=1}^N |P_i| - N + 1)^2$.

Recall that in Step 3, the mass at each s_l is spread out to a set of weighted centroids to obtain \bar{P}^l . For $x_{iq_i}^l \in P_i^l$ for all $i \leq N$ and $c = \sum_{i=1}^N \lambda_i x_{iq_i}^l$ their weighted centroid, we see $\sum_{i=1}^N \lambda_i \|c - x_{iq_i}^l\|^2 \leq \sum_{i=1}^N \lambda_i \|s_l - x_{iq_i}^l\|^2$, independently of how the $x_{iq_i}^l$ are picked from P_i^l . As \bar{P}' is simply the sum of the \bar{P}^l (Step 4), this already implies $\phi(\bar{P}') \leq \phi(\bar{P}_{\text{ex}})$. The algorithm started with a 2-approximation, and thus it is guaranteed to return a \bar{P}' with $\phi(\bar{P}') \leq 2 \cdot \phi(\bar{P})$.

Non-mass split of \bar{P}' is a consequence of two reasons. First, each \bar{P}^l is non-mass splitting by lexicographically maximal choice of the $x_{iq_i}^l$ in Step 3a): Due to this choice, the first weighted centroid c that is constructed will be lexicographically maximal among all (possible) weighted centroids that can be constructed from any x_{iq}^l in the P_i^l . Further, by reducing the mass at each used support point by d_{\min} in Step 3b), at least one of the $d_{iq_i}^l$ becomes 0. The corresponding support point is removed from P_i^l (followed by some reindexing) and thus cannot be used for the construction of weighted centroids in further iterations. Thus the second centroid constructed in the inner loop will be lexicographically strictly smaller than the first one and in particular distinct from it. The same then holds for all subsequent ones. Second, $\text{supp}(\bar{P}^l) \cap \text{supp}(\bar{P}^j) = \emptyset$ for $j \neq l$, because of the lexicographically maximal (d_1, \dots, d_r) : Weighted centroids that would be equally distant from both s_l and s_j cannot appear, because this would have caused a shift of mass to the lower index in Step 2 to resolve such a tie.

The removal of at least one support point from a P_i^l in Step 3b) further implies that $|\bar{P}^l| \leq \sum_{i=1}^N |P_i^l| - N + 1$. Due to $|P_i^l| \leq |P_i|$ and $|\bar{P}_{\text{org}}| \leq \sum_{i=1}^N |P_i| - N + 1$, we obtain the bound

$$|\bar{P}'| = \sum_{l=1}^{|\bar{P}_{\text{org}}|} |\bar{P}^l| \leq \sum_{l=1}^{|\bar{P}_{\text{org}}|} \sum_{i=1}^N |P_i^l| - N + 1 \leq \sum_{l=1}^{|\bar{P}_{\text{org}}|} \sum_{i=1}^N |P_i| - N + 1 \leq \left(\sum_{i=1}^N |P_i| - N + 1 \right)^2.$$

Thus \bar{P}' satisfies all the claimed properties. \square

Next, we prove that Algorithm 2 runs in strongly polynomial time.

Theorem 4 *For all rational input, a 2-approximate non-mass splitting barycenter can be computed in strongly polynomial time.*

Proof. We consider the running time of each part of the algorithm. For readability, we say ‘polynomial’ in this proof in place of ‘strongly polynomial in the bit size of the input’. We use ‘linear’ and ‘quadratic’ to refer to the bit size of the input, too. Note that N , $|P_i|$, and the dimension d are all bounded above by the bit size of the input.

In Step 1, the input for the subsequent steps is created. By sparsity of \bar{P}_{org} , $r \leq \sum_{i=1}^N |P_i| - N + 1$.

For each of the r support points s_l , N images P_i^l with $|P_i^l| \leq |P_i|$ are created. In the application of the rule, each y_{itlk} has to be processed (at most) once. For each y_{itlk} , a single comparison and a fixed number of elementary operations suffices to update the support point and mass in P_i^l . In total, data structures of polynomial size are created in polynomial time.

Step 2 is the processing of (d_1, \dots, d_r) to be lexicographically maximal. For each pair of support points s_l, s_j with $j < l$, we perform the inner part of the loop. Finding q_i in a) can be done by considering all $x_{iq}^l \in P_i^l$ exactly once and comparing the inner products $(s_j - s_l)^T x_{iq}^l$. This is possible in linear time. c is created by the scaling and the sum of N rational d -dimensional vectors.

Part b) begins with the computation of $c - s_j$ and $c - s_l$, then computes $\|c - s_j\|^2 = (c - s_j)^T (c - s_j)$ and $\|c - s_l\|^2 = (c - s_l)^T (c - s_l)$, and then compares the two values. This is possible in quadratic time. Picking the minimal mass among the $x_{iq_i}^l$ is possible in linear time, and so is the update of the masses, the set operations on P_i^l and P_i^j , and the reindexing. By this update, $|P_i^l|$ is reduced by at least one, so the 'go back to a)' statement is followed not more than $|P_i^l|$ times. Summing up, Step 2 runs in polynomial time.

Step 3 performs the spreading of the r support points. Picking a lexicographically maximal support point $x_{iq_i}^l$ in a) can be done by considering all support points in P_i^l once. One saves the current best support point and compares each other support point with respect to their lexicographic order. For identifying the lexicographic order of a pair of d -dimensional support points, (at most) all d of their coefficients have to be compared to each other. This is possible in linear time. Again, c is created by the scaling and the sum of N rational d -dimensional vectors.

In b), we again pick the minimal mass among the $x_{iq_i}^l$ used for the construction of c , which can be done in linear time. The same holds for the update of masses, the set operations on P_i^l , and the reindexing. By this update, $|P_i^l|$ again is reduced by at least one, so the 'go back to a)' statement is followed not more than $|P_i^l|$ times. Further, this implies that $|P^l| \leq \sum_{i=1}^N |P_i^l| - N + 1$. Summing up, the creation of the \bar{P}^l in Step 3 runs in polynomial time.

In Step 4, the \bar{P}^l are summed up to give \bar{P} . This summation is only the creation of a measure with the appropriate mass put on at most $|\bar{P}| \leq (\sum_{i=1}^N |P_i| - N + 1)^2$ support points. Thus all steps run in polynomial time, which proves the claim. \square

4.3 Proofs for 3.3

This section is dedicated to a proof of Theorem 5.

Theorem 5 *Algorithm 3 returns a measure \bar{P}' supported on a subset of S for which $\phi(\bar{P}') \leq 2 \cdot \phi(\bar{P})$, where \bar{P} is a barycenter. Further $|\bar{P}'| \leq \sum_{i=1}^N |P_i| - N + 1$ and \bar{P}' is non-mass splitting.*

Proof. Termination of Algorithm 3 follows because there are only finitely many subsets of S : We know $S' \subset S$ and at the end of Step 2, we update $S_{\text{org}} = S'$ before going back to Step 1. Step 1 computes an optimum over this support S' . Due to $\phi(\bar{P}') \leq \phi(\bar{P}_{\text{org}})$ and termination if $\phi(\bar{P}') = \phi(\bar{P}_{\text{org}})$, for as long as the algorithm keeps running, we have a strictly decreasing sequence of values $\phi(\bar{P}')$. Due to the finite number of subsets of S this can only be a finite sequence. The first approximate barycenter in this sequence is already a 2-approximation and it can only become better throughout the run.

It remains to prove sparsity and the non-mass split property for \bar{P}' . First note that Algorithm 2 will always return a non-mass splitting measure. Further, note that, by Corollary 1, all barycenters \bar{P}_{org} computed in Step 1 have a support that satisfies $|\bar{P}_{\text{org}}| \leq \sum_{i=1}^N |P_i| - N + 1$. So it remains to prove that the output of the final run of Algorithm 2 is a sparse measure \bar{P}' with $|\bar{P}'| \leq \sum_{i=1}^N |P_i| - N + 1$. We do so by showing that in the final run of Algorithm 2, the size of the support of does not increase. We begin by considering Step 3 of Algorithm 2.

Assume P_i^l consists of a single support point x_{i1}^l for all $i \leq N$. Then the unique barycenter \bar{P}^l of the P_i^l is the weighted centroid $c = \sum_{i=1}^N \lambda_i x_{i1}^l$. In this case we can denote the cost of transport

from P^l to all the P_i^l by $\phi(\bar{P}^l) = d_l \cdot \sum_{i=1}^N \lambda_i \|c - x_{i1}^l\|^2$. For all $s \neq c$, we get

$$\phi(\bar{P}^l) = d_l \cdot \sum_{i=1}^N \lambda_i \|c - x_{i1}^l\|^2 < d_l \cdot \sum_{i=1}^N \lambda_i \|s - x_{i1}^l\|^2.$$

Now note that for more general P_i^l , in Step 3 of Algorithm 2, \bar{P}^l is constructed as a set of weighted centroids c of support points x_{iq}^l to which these centroids c transport. These are the ‘building blocks’ of the general \bar{P}^l . Thus

$$\phi(\bar{P}^l) \leq \sum_{i=1}^N \lambda_i \sum_{q=1}^{|P_i^l|} d_{iq}^l \cdot \|s_l - x_{iq}^l\|^2.$$

Informally, it is at least as costly to transport to the measures P_i^l from the support point s_l as from the set of weighted centroids (with appropriate masses) constituting \bar{P}^l . Equality in the above can only hold if the single support point s_l itself is the weighted centroid of P_1^l, \dots, P_N^l that only have a single support point themselves. But this means that transport from s_l is already non-mass splitting and Step 3 of Algorithm 2 just copies s_l with mass d_l to \bar{P}^l .

The algorithm stops when $\phi(\bar{P}') = \phi(\bar{P}_{\text{org}})$. By $\phi(\bar{P}') = \sum_{l=1}^r \phi(\bar{P}^l)$, this means all s_l have to satisfy $\phi(\bar{P}^l) = \sum_{i=1}^N \lambda_i \sum_{q=1}^{|P_i^l|} d_{iq}^l \cdot \|s_l - x_{iq}^l\|^2$. So all s_l are already the weighted centroids of their single-support measures P_i^l .

Note that if a shift of mass from s_l to s_j with $j < l$ had happened in Step 2 of Algorithm 2, then one of the P_i^j would have at least two support points. Thus neither Step 2 nor Step 3 change \bar{P}_{org} in the final run of Algorithm 2 and we obtain $\bar{P}' = \bar{P}_{\text{org}}$. As \bar{P}_{org} satisfies $|\bar{P}_{\text{org}}| \leq \sum_{i=1}^N |P_i| - N + 1$, so does the returned \bar{P}' . This proves sparsity. \square

We close our discussion of Algorithm 3 with a closer look at the approximation error and a worst-case example.

Recall that Algorithm 3 starts with a 2-approximation (in the first run of Step 2) and then improves it iteratively to obtain \bar{P}' . A true barycenter \bar{P} can be rounded to the support $\mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$ by solving the least-squares many-to-one matching

$$\bar{P}_r = \arg \inf_{P_r \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^d)} W_2(P_r, \bar{P})^2. \quad (12)$$

We call \bar{P}_r a ‘rounded’ barycenter. In the following, we distinguish four different measures:

- \bar{P} is an exact barycenter
- \bar{P}_r is a rounded barycenter
- \bar{P}_{org} is an optimal approximate barycenter in $\mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$
- \bar{P}' is the solution of Algorithm 3

By optimality of \bar{P} and \bar{P}_{org} with respect to ϕ in their respective support, we obtain

$$\phi(\bar{P}) \leq \phi(\bar{P}') \leq \phi(\bar{P}_{\text{org}}) \leq \phi(\bar{P}_r).$$

Of course, we are particularly interested in the gap between $\phi(\bar{P})$ and $\phi(\bar{P}')$. Theorem 1 states $\phi(\bar{P}_{\text{org}}) \leq 2 \cdot \phi(\bar{P})$. However, the proof of Theorem 1 actually tells us that $\phi(\bar{P}_r) \leq 2 \cdot \phi(\bar{P})$. Thus the whole sequence of inequalities is bounded by a total approximation factor of 2. This implies that if $\alpha \phi(\bar{P}') = \phi(\bar{P}_{\text{org}})$ for some $\alpha \geq 1$, then $\phi(\bar{P}') \leq \frac{2}{\alpha} \phi(\bar{P})$.

In practice, one obtains a strictly better approximation ratio than 2 for essentially all real-world problems. But there are worst-case examples with $\phi(\bar{P}') = 2\phi(\bar{P})$, which we show in the following example.

Example 3. We revisit Examples 1 and 2, this time giving explicit coordinates for the support points. So let P_1, P_2 be two measures with 3 support points in \mathbb{R}^2 (and $\lambda_1 = \lambda_2 = \frac{1}{2}$), defined as follows: P_1 is supported on $(0, 1)$ with weight $\frac{1}{4}$, $(1, 0)$ with weight $\frac{1}{2}$, and $(2, 1)$ with weight $\frac{1}{4}$. P_2 is supported on $(0, 0)$ with weight $\frac{1}{4}$, $(1, 1)$ with weight $\frac{1}{2}$ and $(2, 0)$ with weight $\frac{1}{4}$. Figure 4 visualizes the example.

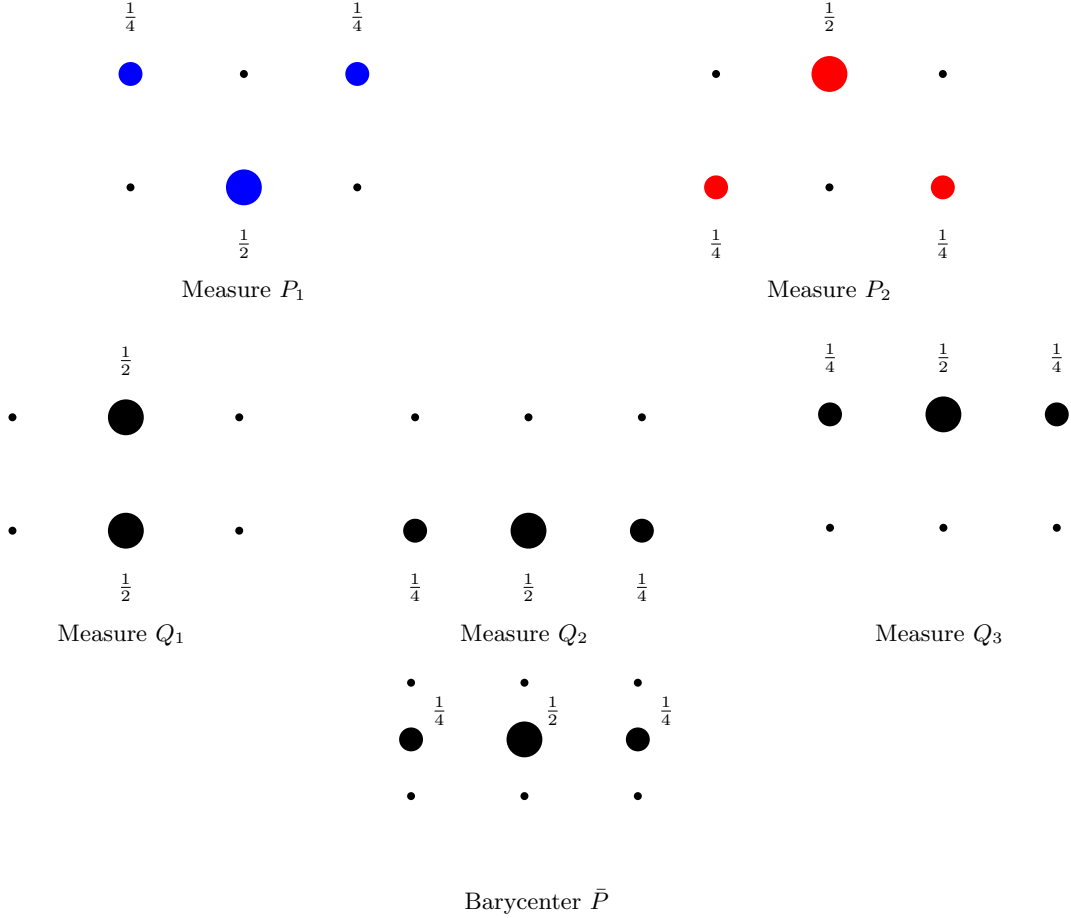


Fig. 4: Two measures P_1, P_2 in the top row. Three measures $Q_1, Q_2, Q_3 \in \mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$ that optimally approximate a barycenter of P_1, P_2 in S_{org} in the second row. The barycenter \bar{P} in the third row, which may round to Q_2 and exhibits $\phi(\bar{P}) = \frac{1}{2}\phi(\bar{P}_r) = \phi(Q_1)$ – while the algorithm starting from Q_1 will just return Q_1 itself.

There are three measures Q_1, Q_2, Q_3 with optimal Wasserstein distances $\phi(\bar{P}_{\text{org}}) = \phi(Q_1) = \phi(Q_2) = \phi(Q_3)$ in $\mathcal{P}_{\text{org}}^2(\mathbb{R}^d)$: Q_1 with weight $\frac{1}{2}$ at both support points $(1, 0)$ and $(1, 1)$, Q_2 with weight $\frac{1}{4}$ at $(0, 0)$, weight $\frac{1}{2}$ at $(1, 0)$, and weight $\frac{1}{4}$ at $(2, 0)$, and Q_3 with weight $\frac{1}{4}$ at $(0, 1)$, weight $\frac{1}{2}$ at $(1, 1)$, and weight $\frac{1}{4}$ at $(2, 1)$. They are depicted in the second row of Figure 4.

A barycenter \bar{P} , depicted in the bottom of the figure, has weight $\frac{1}{4}$ on $(0, \frac{1}{2})$, weight $\frac{1}{2}$ on $(1, \frac{1}{2})$, and weight $\frac{1}{4}$ on $(2, \frac{1}{2})$. For \bar{P} , an optimal solution of Equation (12) is either $\bar{P}_r = Q_2$ or $\bar{P}_r = Q_3$. In both cases, $\phi(\bar{P}) = \frac{1}{2}\phi(\bar{P}_r) = \phi(\bar{P}_{\text{org}})$ and we obtain a configuration with

$$\phi(\bar{P}) = \frac{1}{2}\phi(\bar{P}_0) = \frac{1}{2}\phi(P_{\text{org}}) = \frac{1}{2}\phi(\bar{P}_r).$$

Informally, the algorithm is not guaranteed to improve on the initial 2-approximation, but only if the initial approximate barycenter already was non-mass splitting. \square

5 Practical Computations

We implemented Algorithm 3 in the Julia language ([4,17]) using Clp as linear programming solver. To keep the number of iterations low, we implemented Step 3 of Algorithm 2 as the exact computation of a barycenter \bar{P}^l when the number of support points to which a given s_l transports is low.

Let us exhibit some data from some sample computations for the widely-used MNIST database of handwritten digits [16]. We also used it for the example in Figure 1. As input, we chose the four digits representing number six depicted in Figure 5. They have a barycenter depicted in the bottom of the figure (for all $\lambda_i = \frac{1}{4}$). The computation of this barycenter took roughly 120 seconds on a standard laptop.

In Figure 6 we exhibit a run of Algorithm 3 for the same input. It completed in about 10 seconds. For larger data sets, the difference between the running times will be more extreme due to the exponential size of the LP for the computation of the exact barycenter.

Each row shows one of the iterations. The approximation of the barycenter in the original support is already a 1.142-approximation of the true barycenter (top left), i.e. $\phi(\bar{P}_{\text{org}}) \leq 1.142 \cdot \phi(\bar{P})$, which we denote as an additive 14.2%-error in the figure. The first split-up using Algorithm 2 gives an improvement to a 4.3% error (top right). This is further improved on to a 2.0% by computing an optimum in the support of the previous approximation (bottom left). In this run, now all the support points are already non-mass splitting and they are the weighted centroids of the support points to which they transport, so the algorithm terminates.

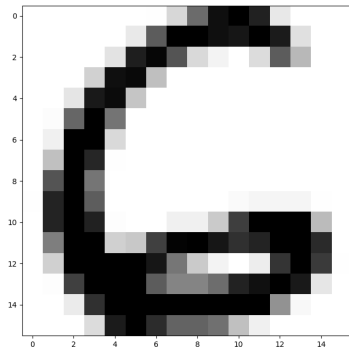
This early termination is not surprising, there are several reasons: The small data set, the low initial error, and most importantly, the implementation of an exact barycenter computation to find (globally) optimal P^l for all support points s_l . We observed a similar behavior for larger computations (for 20 measures), where approximation errors in the original support are already low (much lower than the guaranteed bound of 2), and only the first three to four iterations provided improvements that were distinguishable from numerical errors. For practical applications in which only the approximation error and a short running time is relevant, we recommend a hardcoded termination after 2 to 3 iterations.

Acknowledgments

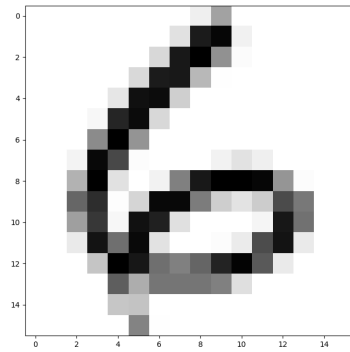
We would like to thank Ethan Anderes for the support implementing our algorithms in the Julia language and Jacob Miller for the helpful discussions.

References

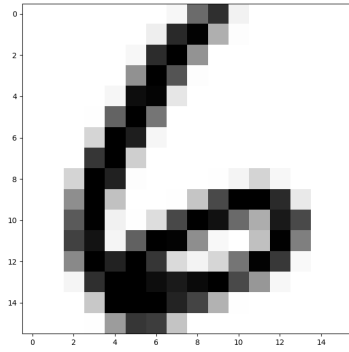
1. M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43 (2):904–924, 2011.
2. E. Anderes, S. Borgwardt, and J. Miller. Discrete Wasserstein Barycenters: Optimal Transport for Discrete Data. *Mathematical Methods of Operations Research*, 84 (2):389–409, 2016.
3. M. Beiglböck, P. Henry-Labordere, and F. Penkner. Model-independent bounds for option prices – a mass transport approach. *Finance and Stochastics*, 17 (3):477–501, 2013.
4. J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *CoRR*, abs/1411.1607, 2014.
5. J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *eprint arXiv:1212.2562*, 2012.
6. E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution’s template estimate with Wasserstein metrics. *Bernoulli*, 21 (2):740–759, 2015.
7. G. Buttazzo, L. De Pascale, and P. Gori-Giorgi. Optimal-transport formulation of electronic density-functional theory. *Phys. Rev. A*, 85:062502, 2012.



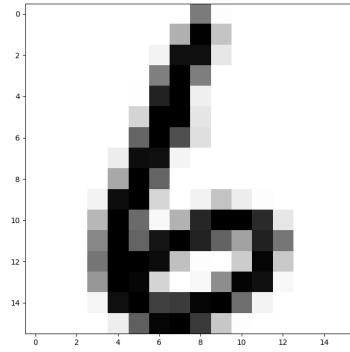
Measure P_1



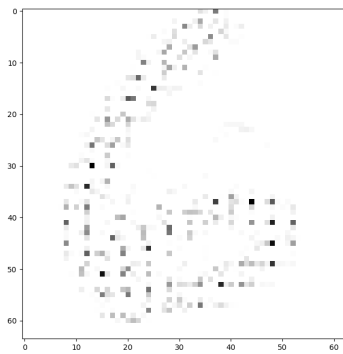
Measure P_2



Measure P_3

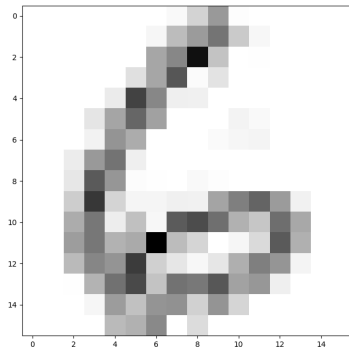


Measure P_4

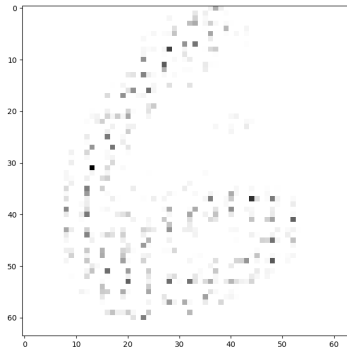


Barycenter \bar{P}

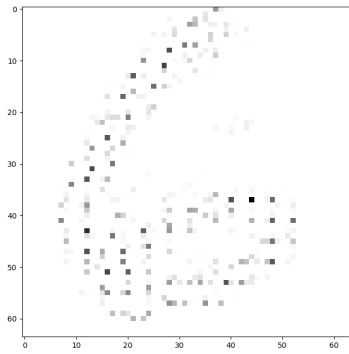
Fig. 5: Four measures P_1, \dots, P_4 , scans of handwritten digits six, supported on a 16×16 grid. The barycenter \bar{P} in the bottom row.



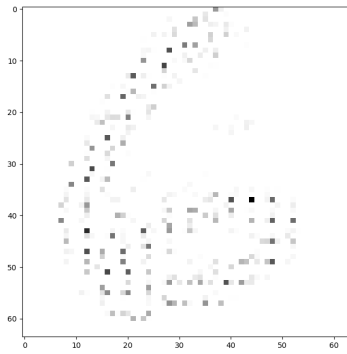
Iteration 1, Step 1, Error 14.2%



Iteration 1, Step 2, Error 4.3%



Iteration 2, Step 1, Error 2.0%



Iteration 2, Step 2, termination

Fig. 6: Stages of a run of Algorithm 3. The algorithm already terminates after 2 iterations.

8. G. Carlier and I. Ekeland. Matching for teams. *Economic Theory*, 42 (2):397–418, 2010.
9. G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49 (6):1621–1642, 2015.
10. P.-A. Chiaporri, R. McCann, and L. Nesheim. Hedonic price equilibria, stable matching and optimal transport; equivalence, topology and uniqueness. *Economic Theory*, 42 (2):317–354, 2010.
11. C. Cotar, G. Friesecke, and C. Klüppelberg. Density functional theory and optimal transportation with coulomb cost. *Communications on Pure and Applied Mathematics*, 66 (4):548–599, 2013.
12. M. Cuturi and A. Doucet. Fast Computation of Wasserstein Barycenters. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 685–693. JMLR Workshop and Conference Proceedings, 2014.
13. E. del Barrio, J.A. Cuesta-Albertos, C. Matrán, and A. Mayo-Íscar. Robust clustering tools based on optimal transportation. *eprint arXiv:1607.01179*, 2016.
14. A. Galichon, P. Henry-Labordere, and N. Touzi. A stochastic control approach to non-arbitrage bounds given marginals, with an application to lookback options. *Annals of Applied Probability*, 24 (1):312–336, 2014.
15. A. Jain, Y. Zhong, and M.-P. Dubuisson-Jolly. Deformable template models: A review. *Signal Processing*, 71 (2):109–129, 1998.
16. Y. LeCu, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
17. M. Lubin and I. Dunning. Computing in operations research using julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.
18. Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12), 2011.
19. J. Miller. Transportation Networks and Matroids: Algorithms through Circuits and Polyhedrality, 2016. Ph.D. thesis, University of California Davis.
20. E. Munch, K. Turner, P. Bendich, S. Mukherjee, J. Mattingly, and J. Harer. Probabilistic frechet means for time varying persistence diagrams. *Electronic Journal of Statistics*, 9:1173–1204, 2015.
21. B. Pass. On the local structure of optimal measures in the multi-marginal optimal transportation problem. *Calculus of Variations and Partial Differential Equations*, 43 (3-4):529–536, 2011.
22. B. Pass. Uniqueness and Monge Solutions in the Multimarginal Optimal Transportation Problem. *SIAM Journal on Mathematical Analysis*, 43 (6):2758–2775, 2011.
23. B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264 (4):947–963, 2013.
24. B. Pass. Multi-marginal optimal transport and multi-agent matching problems: Uniqueness and structure of solutions. *Discrete and Continuous Dynamical Systems A*, 34 (4):1623–1639, 2014.
25. J. Rabin, G. Peyre, J. Delon, and M. Bernot. Wasserstein Barycenter and its Application to Texture Mixing. *Scale Space and Variational Methods in Computer Vision. Lecture Notes in Computer Science*, 6667:435–446, 2012.
26. E. Tardos. A strongly polynomial algorithm to solve combinatorial linear programs. *Operations Research*, 34(2):250–256, 1986.
27. A. Trounev and L. Younes. Local Geometry of Deformable Templates. *SIAM Journal on Mathematical Analysis*, 37 (1):17–59, 2005.
28. K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Frechet means for distributions of persistence diagrams. *Discrete and Computational Geometry*, 52(1):44–70, 2014.
29. C. Villani. *Topics in Optimal Transportation*, volume 58. 2003.
30. C. Villani. *Optimal transport: old and new*, volume 338. 2009.
31. J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, 2017.
32. Y. Zemel and V. Panaretos. Fréchet Means and Procrustes Analysis in Wasserstein Space. *eprint arXiv:1701.06876*, 2017.